UNITED STATES UTILITY PATENT APPLICATION FOR

NONLINEAR MODELING OF GENE NETWORKS FROM TIME SERIES GENE EXPRESSION DATA

Inventors: Satoru Miyano
Seiya Imoto
Sun Yong Kim

CERTIFICATE OF MAILING BY "EXPRESS MAIL" UNDER 37 C.F.R. § 1.10

"Express Mail" mailing label number: EV 326 482 189 US
Date of Mailing: November 18, 2003

I hereby certify that this correspondence is being deposited with the United States Postal Service, utilizing the "Express Mail Post Office to Addressee" service addressed to **BOX UTILITY PATENT APPLICATION, Assistant Commissioner for Patents, Washington, D.C. 20231** and mailed on the above Date of Mailing with the above "Express Mail" mailing label number.

D. Benjamin Borson (Signature)
D. Benjamin Borson, Reg. No. 42,349
Signature Date: November 18, 2003

NONLINEAR MODELING OF GENE NETWORKS FROM TIME SERIES GENE EXPRESSION DATA

Related Application:

This application claims priority under 35 U.S.C. §119(e) to United States Provisional Application Serial No: 60/427,448 filed November 19, 2002. This application is herein incorporated fully by reference.

Field of the Invention

This invention relates to the use of Bayesian models with nonparametric regression to infer network relationships between genes from time series studies of gene expression. In particular, the invention relates to methods involving minimizing a criterion, $BNRC_{dynamic}$ to infer optimal network relationships.

BACKGROUND

One of the most important aspects of current research and development in the life sciences, medicine, drug discovery and development and pharmaceutical industries is the need to develop methods and devices for interpreting large amounts of raw data and drawing conclusions based on such data. Bioinformatics has contributed substantially to the understanding of systems biology and promises to produce even greater understanding of the complex relationships between components of living systems. In particular, with the advent of new methods for rapidly detecting expressed genes and for quantifying expression of genes, bioinformatics can be used to predict potential therapeutic targets even without knowing with certainty, the exact roles a particular gene(s) may play in the biology of an organism.

Simulation of genetic systems is a central topic of systems biology. Because simulations can be based on biological knowledge, a network estimation method can support biological simulation by predicting or inferring previously unknown relationships.

In particular, development of microarray technology has permitted studies of expression of a large number of genes from a variety of organisms. A very large amount of raw data can be obtained from a number of genes from an organism, and gene expression can be studied by intervention by either mutation, disease or drugs. Finding that a particular gene's expression is increased in a particular disease or in response to a particular intervention may lead one to believe that that gene is directly involved in the disease process or drug response. However, in biological organisms genes rarely are independently regulated by any such intervention, in that many genes can be affected by a particular intervention. Because a large number of different genes may be so affected, understanding the cause and effect relationships between genes in such studies is very difficult. Thus, much effort is being expended to develop methods for determining cause and effect relationships between genes, which genes are central to a biological phenomenon, and which genes' expression(s) are peripheral to the biological process under study. Although such peripheral gene's expression may be useful as a marker of a biological or pathophysiological condition, if such a gene is not central to physiological or pathophysiological conditions, developing drugs based on such genes may not be worth the efforts. In contrast, for genes identified to be central to a process, development of drugs or other interventions may be crucial to developing treatments for conditions associated with altered expression of genes.

Development of Bayesian network analysis for estimating a gene network from microarray gene expression data has received considerable attention and many successful investigations have been reported (Friedman et al [13]; Imoto et al [14]; Pe'er et al. [18] and our own work [U.S. Patent Application Serial No: 10/259,723 herein incorporated fully by reference].

However, a shortcoming of traditional Bayesian network models is that they cannot construct cyclic networks, while certain real gene regulatory mechanisms have cyclic components. Recently, a dynamic Bayesian network model (Bilmes et al. [3]; Friedman et al [12]; Someren et al [19] has been propsed for constructing a gene network with cyclic regulatory components. Dynamic Bayesian network is based on time series data, and usually the data can be discritized into several classes. Thus, a dynamic network model can depend on the setting of the thresholds for

the discritizing process, and unfortunately, the discritization can lead to loss of information. Imoto et al. [14, 15] proposed a network estimation method based on a Bayesian network and nonparametric regression for a solution to avoid discretization and for capturing non-linear relations among genes. However, Bayesian networks and nonparametric regression models [14, 15] still may not adequately solve networks having cyclic regulatory components.

SUMMARY

In certain embodiments, this invention includes the use of time-series expression data in a Bayesian network model with nonparametric regression. Using time series expression data, we can identify cyclic regulatory components. In other embodiments, time delay information can be incorporated into a Bayesian/nonparametric regression model, which then can extract even nonlinear relations among genes. In certain of these embodiments, an ordinal differential equation model can be used as an alternative. We also have developed new criteria for choosing an optimal network from a Bayesian statistical point of view. Such criteria can optimize a network structure based on data having noise.

BRIEF DESCRIPTION OF THE FIGURES

This invention is described with reference to specific embodiments thereof. Additional aspects of the invention are found the Examples and in the Figures, in which:

Figure 1 depicts a schematic illustration of time dynamics in gene expression.

Figures 2a and 2b depict diagrams of network relationships of genes involved in cell cycle regulation in yeast, compiled in KEGG.

Figure 2a depicts genes in cyclin-dependent protein kinase pathways.

Figure 2b depicts network relationships between genes described in Figure 2a involved in regulating cyclin-dependent protein kinases.

Figures 3a - 3c depict diagrams of network relationships of yeast genes involved in metabolic pathways.

Figure 3a depicts several genes involved in metabolic pathways.

Figure 3b depicts network relationships between genes described in Figure 3a derived from a Bayesian/nonparametric regression model.

Figure 3c depicts network relationships between genes described in Figure 3a derived from a dynamic Bayesian/nonparametric regression model.

DETAILED DESCRIPTION

In general, a dynamic Bayesian network model can be obtained using any suitable method for determining gene expression. In certain embodiments, microarray experiments are desirable because a large number of genes can be studied from a single sample applied to the array, making relative differences in gene expression easy to determine. It may be desirable to improve accuracy of microarray methods by subtracting background signals from the signal reflecting true gene expression and/or correcting for inherent differences in labels used to measure gene expression (e.g., cy3/cy5)

Using a Bayesian network framework, we consider a gene as a random variable and decompose the joint probability into the product of conditional probabilities. For example, if we have a series of observations of the random vector, we can denote the probability of obtaining a given observation can depend upon the conditional probability densities. In certain embodiments, one can use nonparametric regression models for capturing the relationships between the variables. A variety of graphic tools can be used to elucidate the relationships. For example, polynomials, Fourier series, regression spline gases, B-spline bases, wavelet bases and the like can be used for defining a graph of gene relationships. Certain methods to elucidate network relationships are disclosed in U.S. Patent Application Serial No: 10/259,723, herein incorporated fully by reference. One difficulty in selecting a proper graph is to properly evaluate variance and noise in the system.

In some embodiments of this invention, networks can be constructed using Bayesian estimation with nonparametric regression using data from time series studies. In many gene

networks, an intervention leads to alteration in expression of certain genes before alterations in other genes are observed. One may infer that expression of certain genes after an intervention that occur later in time, may be causally related to genes whose expression is early. Time series information is useful to define "early" or genes and "late" or gene. It is unlikely that an alteration in expression of a late gene could be a cause of an alteration of expression of an early gene, whose expression is altered sooner in time than that of a late gene. Although this presumption may not apply in all cases, it is more probable that early genes are more likely "upstream" in a network than are late genes, which are more likely to be "downstream" genes. Therefore, time relations of gene expression can be useful to modify Bayesian estimation and nonparametric regression to provide a more reliable network solution.

In aspects of this invention, we extend the Bayesian network and nonparametric regression model to a dynamic Bayesian network model, which can be used to construct cyclic relationships when one has time series gene expression data. Information on time delay between changes in gene expression can be included in a model easily, and the model can extract even nonlinear relations among genes easily.

In certain embodiments, for constructing a gene network with cyclic regulatory components, an ordinal differential equation model (Chen et al. [5]; de Hoon et al. [8] can be used. However, this model is based on a linear system and may be unsuitable for capturing complex phenomena. We have derived a new criterion for choosing an optimal network from Bayesian statistical point of view [2]. The criterion can optimize network structure, which gives the best representation of the gene interactions described by the data with noise. The new criterion is herein termed $\mathrm{BNRC}_{dynamic}$.

$\mathrm{BNRC}_{dynamic}$ can be evaluated using a first-order Markov relation as illustrated in Figure 1. In such a relationship, an upstream gene X_i is depicted as having an effect (right arrow) on one or more downstream genes X_2, which has an effect on X_3 (not shown), and so on, until an effect on X_n is observed. In situations in which X_1 has no "upstream" gene of its own, X_1 is termed a "parent" gene within the network. Genes under influence of a parent gene are termed "target"

genes. Note that the use of "target gene" in this context is not to be confused with a gene that is a target for intervention, such as by a potential drug. In fact, parent genes may be targets for therapeutic intervention. Under this scheme, an effect on X_n cannot be observed until effects on X_1, X_2, etc. have been elicited. Note that Figure 1 illustrates a "series" cause/effect relationship, without parallel or feedback systems are present, whereas in many genetic systems, there are series effects, and "parallel" effects, in which two or more genes can either be affected by an upstream gene, and/or can themselves affect a downstream gene. Moreover, circular effects ("feedback") can be present, in which a gene X_a can affect another gene X_b, which can affect X_c, which itself can affect X_a (or X_b). Moreover, such feedback may be either positive, in which X_c stimulates X_a or "negative" in which X_c inhibits X_a. Further complexities can arise in situations in which both series, parallel, positive feedback and negative feedback relationships are present.

In general, relationships between time points may be arbitrary, but in some cases it can be advantageous to use pre-selected time points based on knowledge of the biological effects of the genes and their expression dynamics under study. Under first order conditions, a joint probability can be decomposed as shown in equation (1) in Example 1 below. A conditional probability can then be decomposed into the product of conditional probabilities using equation (2) in Example 1. Equations (1) and (2) can hold and the density function can be used instead of a probability measure. Therefore, the dynamic Bayesian network can be represented, for example, using densities described in Example 1 to arrive at the local network structure of a gene and its parent genes according to equation (3) in Example 1.

A dynamic Bayesian model with nonparametric regression can be applied, for example, as described in Example 2. Once experimental data is collected, a the solution to the network can be considered to be a statistical model selection problem. In certain embodiments, we can solve this problem using Bayesian approach and derive a criterion for evaluating the goodness of the dynamic Bayesian network and nonparametric regression methods. Assuming a prior distribution, marginal likelihood and posterior probability can be determined according to equation (4) in Example 2. Subsequent construction of a genetic network involves computation of a high

dimensional integral as depicted in equation (4). In some embodiments, a Laplace method for integrals, for example, can be used to approximate the integral. Therefore, the criterion $BNRC_{dynamic}$ as shown in equation (5) in Example 2 can be solved.

To apply $BNRC_{dynamic}$ to an experimental system, cDNA microarray data, for example, can be obtained experimentally at a number of time points after affecting the genetic system. To smooth curves, we can use spline functions, for example B-splines as depicted in Example 3. $BNRC_{dynamic}$ can be decomposed according to equation (6) in Example 3. Optimal network relationships are obtained when $BNRC_{dynamic}$ is minimized.

Using dynamic Bayesian network models with nonparametric regression and the criterion $BNRC_{dynamic}$, we can formulate a network learning process. However, determining which genes are parent genes and which are target genes can be time consuming when all possible gene combinations and relationships are considered. To reduce the number of analyses needed, we can select candidate parent genes. Subsequently a greedy hill-climbing algorithm can be used. $BNRC_{dynamic}$ is calculated and then an addition parent gene is either added or deleted, and $BNRC_{dynamic}$ is re-evaluated according to Step 2 in Example 3. The process is repeated until an appropriate convergence is found. Then, the order of computation is permutated and $BNRC_{dynamic}$ is reevaluated. The optimal network give the smallest $BNRC_{dynamic}$.

A specific illustration of the above methods are shown in Example 4 in Figures 2a and 2b. The efficiencies of the methods are shown through analysis of gene expression data from *Saccharomyces cerevisiae*. Figure 2a depicts a group of *S. cerevisiae* genes involved in regulation of cell cycle. The genes are depicted as grouped based in the overall metabolic pathways involved and focus on the cyclin-dependent protein kinase gene (YBR160w). Note that the parent/target gene network relationships are unknown based on Figure 2a. In contrast, using methods of this invention, network relationships of those genes can be evaluated and are depicted in Figure 2b.

Another example is depicted in Figures 3a - 3c. Figure 3a depicts genes involved in metabolic pathways. Figure 3a shows no gene network relationships. Figure 3b depicts a network

solution obtained using Bayesian network analysis with nonparametric regression, but without consideration of $BNRC_{dynamic}$. Figure 3c depicts a network solution obtained by minimizing $BNRC_{dynamic}$. Note that in Figure 3b, the network relationships are simpler, and compared to those depicted in Figure 3b, there are many fewer false positive relationships ("x").

Boundaries between groups of genes in a network can be determined using methods known in the art, for example, bootstrap methods. Such methods include determining the intensity of an edge
using the following steps.

(1) providing a bootstrap gene expression matrix by randomly sampling a number of times, with replacement, from the original gene library expression data;

(2) estimating the genetic network for $gene_i$ and $gene_j$;

(3) repeating steps (1) and (2) *T* times, thereby producing *T* genetic networks; and

(4) calculating the bootstrap edge intensity between $gene_i$ and $gene_j$ as $(t_1 + t_2)/T$.

Advantages of the new methods compared with other network estimation methods such as Bayesian and Boolean Networks include: (1) time information can be incorporated easily; (2) microarray data can be analyzed as continuous data without extra data pre-treatments such as discretization; and (3) fewer false positive relationships are found. Even nonlinear relations can be detected and modeled by embodiments of this invention. Methods of this invention are useful for analyzing genetic networks and for development of new pharmaceuticals which target particularly genes that control genetic expression of important genes. Thus, methods of this invention can decrease the time needed to identify drug targets and therefore can decrease the time needed to develop new treatments.

Other aspects of methods of this invention are described in the Examples below.

EXAMPLES

The examples presented below represent specific embodiments of this invention. Other

aspects of the invention can be developed by persons of ordinary skill in the art without undue experimentation. All such embodiments are considered part of this invention.

Example 1: Bayesian Network and Nonparametric Regression

Suppose that we have an $n \times p$ microarray gene expression data matrix $\boldsymbol{X}$, where n and p are the numbers of microarrays and genes, respectively. Usually, the number of genes p is much larger than the number of microarrays, n. In the estimation of a gene network based on the Bayesian network, a gene is considered as a random variable. When we model a gene network by using statistical models described by the density or probability function, the statistical model should include p random variables. However, we have only n samples and n is usually much smaller than p. In such case, the inference of the model is quite difficult or impossible, because the model has many parameters and the number of samples is not enough for estimating the parameters. The Bayesian network model has been advocated in such modeling.

In the context of the dynamic Bayesian network, we consider the time series data and the ith column vector $\boldsymbol{x}_i$ of $\boldsymbol{X}$ corresponds to the states of p genes at time i. As for the time dependency, we consider the first order Markov relation described in Figure 1. Under this condition, the joint probability can be

decomposed as follows:

$$P(X_{11},\dots,X_{np}) = P(\boldsymbol{X}_1)P(\boldsymbol{X}_2|\boldsymbol{X}_1)\times\cdots\times P(\boldsymbol{X}_n|\boldsymbol{X}_{n-1}), \tag{1}$$

where $\boldsymbol{X}_i = (X_{i1},\dots,X_{ip})$ is a random variable vector of p genes at time i. The conditional probability $P(\boldsymbol{X}_i|\boldsymbol{X}_{i-1})$ can also be decomposed into the product of conditional probabilities of the form

$$P(\boldsymbol{X}_i|\boldsymbol{X}_{i-1}) = P(X_{i1}|\boldsymbol{P}_{i-1,1})\times\cdots\times P(X_{ip}|\boldsymbol{P}_{i-1,p}), \tag{2}$$

where $\boldsymbol{P}_{i-1,j}$ is the state vector of the parent genes of jth gene at time $i-1$. The equations (1) and (2) hold when we use the density function instead of the probability measure. Hence, the dynamic Bayesian network can then be represented by using densities as follows:

$$\begin{aligned} f(x_{11},\dots,x_{np}) &= f_1(\boldsymbol{x}_1)f_2(\boldsymbol{x}_2|\boldsymbol{x}_1)\times\cdots\times f_n(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}) \\ &= f_1(\boldsymbol{x}_1)\prod_{i=2}^{n} g_1(x_{i1}|\boldsymbol{p}_{i-1,1})\times\cdots\times g_p(x_{ip}|\boldsymbol{p}_{i-1,p}) \\ &= f_1(\boldsymbol{x}_1)\prod_{j=1}^{p}\left\{\prod_{i=2}^{n} g_j(x_{ij}|\boldsymbol{p}_{i-1,j})\right\}. \end{aligned}$$

Here we have the decomposition from (2)

$$f_i(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}) = g_1(x_{i1}|\boldsymbol{p}_{i-1,1})\times\cdots\times g_p(x_{ip}|\boldsymbol{p}_{i-1,p}),$$

where $\boldsymbol{p}_{i-1,j} = (p^{(j)}_{i-1,1},\dots,p^{(j)}_{i-1,q_j})$ is a q_j-dimensional observation vector of parent genes.

For modeling the relationship between x_{ij} and $p_{i-1,j}$, we use the nonparametric additive regression model as follows:

$$x_{ij} = m_{j1}(p_{i-1,1}^{(j)}) + \cdots + m_{jq_j}(p_{i-1,q_j}^{(j)}) + \varepsilon_{ij},$$

where ε_{ij} depends independently and normally on mean 0 and variance σ_j^2. Here, $m_{jk}(\cdot)$ is a smooth function from R to R and can be expressed by using the linear combination of basis functions

$$m_{jk}(p_{i-1,k}^{(j)}) = \sum_{m=1}^{M_{jk}} \gamma_{mk}^{(j)} b_{mk}^{(j)}(p_{i-1,k}^{(j)}), \quad k = 1, \ldots, q_j,$$

where $\gamma_{1k}^{(j)}, \ldots, \gamma_{M_{jk}k}^{(j)}$ are unknown coefficient parameters and $\{b_{1k}^{(j)}(\cdot), \ldots, b_{M_{jk}k}^{(j)}(\cdot)\}$ is the prescribed set of basis functions. Then we define a dynamic Bayesian network and nonparametric regression model of the form

$$\begin{aligned} &f(x_{11}, \ldots, x_{np}; \theta_G) \\ &\quad = f_1(x_1) \prod_{j=1}^{p} \left[\prod_{i=2}^{n} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{ -\frac{(x_{ij} - \mu(p_{i-1,j}))^2}{2\sigma_j^2} \right\} \right], \end{aligned}$$

where $\mu(p_{i-1,j}) = m_{j1}(p_{i-1,1}^{(j)}) + \cdots + m_{jq_j}(p_{i-1,q_j}^{(j)})$. When jth gene has no parent genes, $\mu(p_{i-1,j})$ is resulted in the constant μ_j.

We assume $f_1(x_1) = g_1(x_{11}) \times \cdots \times g_1(x_{1p})$ and the joint density $f(x_{11}, \ldots, x_{np}; \theta_G)$ can then be rewritten as

$$\begin{aligned} f(x_{11}, \ldots, x_{np}; \theta_G) &= \prod_{j=1}^{p} \left[g_1(x_{1j}) \prod_{i=2}^{n} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{ -\frac{(x_{ij} - \mu(p_{i-1,j}))^2}{2\sigma_j^2} \right\} \right] \\ &= \prod_{j=1}^{p} \prod_{i=1}^{n} g_j(x_{ij} | p_{i-1,j}; \theta_j), \qquad (3) \end{aligned}$$

where $p_{0j} = \emptyset$. Thus, $g_j(x_{ij} | p_{i-1,j}; \theta_j)$ represents the local structure of jth gene and its parent genes.

Example 2: Derivation of a Criterion for Selecting a Network

The dynamic Bayesian network and nonparametric regression model introduced in the previous section can be constructed when we fix the network structure and estimated by a suitable procedure. However, the gene network is generally unknown and we should estimate an optimal network based on the data. This problem can be viewed as a statistical model selection problem (see e.g., Akaike [1]; Konishi and Kitagawa [17]; Burnham and Anderson [4]; Konishi [16]). We solve this problem from the Bayesian statistical approach and derive a criterion for evaluating the goodness of the dynamic Bayesian network and nonparametric regression model.

Let $\pi(\theta_G|\lambda)$ be a prior distribution on the parameter θ_G in the dynamic Bayesian network and nonparametric regression model and let $\log \pi(\theta_G|\lambda) = O(n)$. The marginal likelihood can be represented as

$$\int f(x_{11}, \ldots, x_{np}; \theta_G)\pi(\theta_G|\lambda)d\theta_G.$$

Thus, when the data is given, the posterior probability of the network G is

$$\pi_{post}(G|X) = \frac{\pi_{prior}(G)\int f(x_{11}, \ldots, x_{np}; \theta_G)\pi(\theta_G|\lambda)d\theta_G}{\sum_G \left\{\pi_{prior}(G)\int f(x_{11}, \ldots, x_{np}; \theta_G)\pi(\theta_G|\lambda)d\theta_G\right\}}, \qquad (4)$$

where $\pi_{prior}(G)$ is the prior probability of the network G. The denominator of (4) does not relate to model evaluation. Therefore, the evaluation of the network depends on the magnitude of numerator. Hence, we can choose an optimal network as the maximizer of

$$\pi_{prior}(G)\int f(x_{11}, \ldots, x_{np}; \theta_G)\pi(\theta_G|\lambda)d\theta_G.$$

It is clear that the essential point for constructing a network selection criterion is how to compute the high dimensional integral. Imoto *et al.* [14, 15] used the Laplace approximation for integrals (see also Tinerey and Kadane [21]; Davison [6]) and we can apply this technique to the dynamic Bayesian network model and nonparametric regression directly. Hence, we have a criterion, named $\mathrm{BNRC}_{dynamic}$, of the form

$$
\begin{aligned}
&\mathrm{BNRC}_{dynamic}(G) \\
&\quad = -2\log\left\{\pi_{prior}(G)\int f(x_{11},...,x_{np};\boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})d\boldsymbol{\theta}_G\right\} \\
&\quad \approx -2\log\pi_{prior}(G) - r\log(2\pi/n) + \log|J_\lambda(\hat{\boldsymbol{\theta}}_G)| - 2nl_\lambda(\hat{\boldsymbol{\theta}}_G|\boldsymbol{X}_n),
\end{aligned}
\qquad (5)
$$

where r is the dimension of $\boldsymbol{\theta}_G$,

$$
\begin{aligned}
l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X}_n) &= \log f(x_{11},...,x_{np};\boldsymbol{\theta}_G)/n + \log\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})/n, \\
J_\lambda(\boldsymbol{\theta}_G) &= -\partial^2\{l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X}_n)\}/\partial\boldsymbol{\theta}_G\partial\boldsymbol{\theta}_G^T
\end{aligned}
$$

and $\hat{\boldsymbol{\theta}}_G$ is the mode of $l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X}_n)$. The optimal graph is chosen such that the criterion $\mathrm{BNRC}_{dynamic}$ (5) is minimal.

Example 3: Estimation of a Gene Network

In this section, we show a concrete strategy for estimating a gene network from cDNA microarray time series gene expression data.

3.1 Nonparametric Regression

We use the basis function approach for constructing the smooth function $m_{jk}(\cdot)$ described in Section 2. In this paper we use B-splines (de Boor [7]) as the basis functions. De Boor's algorithm (de Boor [7], Chapter 10, p.130 (3)) is a useful method for computing B-splines of any degree. We use 20 B-splines with equidistance knots (see also, Dierckx [10]; Eiler and Marx [11] for the details of B-spline).

3.2 Prior Distribution on the Parameter in the Model

For the prior distribution on the parameter $\boldsymbol{\theta}_G$, suppose that the parameter vectors $\boldsymbol{\theta}_j$ are independent one another, the prior distribution can then be decomposed as $\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda}) = \prod_{j=1}^{p} \pi_j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j)$. Suppose that the prior distribution $\pi_j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j)$ is factorized as $\pi_j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j) = \prod_{k=1}^{q_j} \pi_{jk}(\boldsymbol{\gamma}_{jk}|\lambda_{jk})$, where λ_{jk} are hyper parameters. We use a singular M_{jk} variate normal distribution as the prior distribution on $\boldsymbol{\gamma}_{jk}$,

$$\pi_{jk}(\boldsymbol{\gamma}_{jk}|\lambda_{jk}) = \left(\frac{2\pi}{n\lambda_{jk}}\right)^{-(M_{jk}-2)/2} |K_{jk}|_{+}^{1/2} \exp\left(-\frac{n\lambda_{jk}}{2}\boldsymbol{\gamma}_{jk}^T K_{jk}\boldsymbol{\gamma}_{jk}\right),$$

where K_{jk} is an $M_{jk} \times M_{jk}$ symmetric positive semidefinite matrix satisfying $\gamma_{jk}^T K_{jk} \gamma_{jk} = \sum_{\alpha=3}^{M_{jk}} (\gamma_{\alpha k}^{(j)} - 2\gamma_{\alpha-1,k}^{(j)} + \gamma_{\alpha-2,k}^{(j)})^2$. This setting of the prior distribution on θ_G is the same as Imoto *et al.* [14, 15] and the details are in those papers.

3.3 Proposed Criterion

By using the prior distributions in Section 4.2, the $\mathrm{BNRC}_{dynamic}$ can be decomposed as follows:

$$\mathrm{BNRC}_{dynamic} = \sum_{j=1}^{p} \mathrm{BNRC}_{dynamic}^{(j)}, \tag{6}$$

where $\mathrm{BNRC}_{dynamic}^{(j)}$ is a local criterion score of jth gene and is defined by

$$\begin{aligned}
&\mathrm{BNRC}_{dynamic}^{(j)} \\
&\quad = -2\log\left\{\int \pi_{prior}(L_j)\prod_{i=1}^{n} g_j(x_{ij}|\boldsymbol{p}_{i-1,j};\boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j)d\boldsymbol{\theta}_j\right\} \\
&\quad \approx -2\log\pi_{prior}(L_j) - r_j\log(2\pi/n) + \log|J_{\lambda_j}^{(j)}(\hat{\boldsymbol{\theta}}_j)| - 2nl_{\lambda_j}^{(j)}(\hat{\boldsymbol{\theta}}_j|\boldsymbol{X}),
\end{aligned}$$

where r_j is the dimension of $\boldsymbol{\theta}_j$,

$$\begin{aligned}
l_{\lambda_j}^{(j)}(\hat{\boldsymbol{\theta}}_j|\boldsymbol{X}) &= \sum_{i=1}^{n}\log g_j(x_{ij}|\boldsymbol{p}_{i-1,j};\boldsymbol{\theta}_j)/n + \log\pi(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j)/n, \\
J_{\lambda_j}^{(j)}(\hat{\boldsymbol{\theta}}_j) &= -\partial^2\{l_{\lambda_j}^{(j)}(\hat{\boldsymbol{\theta}}_j|\boldsymbol{X})\}/\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_j^T
\end{aligned}$$

and $\hat{\boldsymbol{\theta}}_j$ is the mode of $l_{\lambda_j}^{(j)}(\boldsymbol{\theta}_j|\boldsymbol{X})$. Here $\pi_{prior}(L_j)$ are prior probabilities satisfying $\sum_{j=1}^{p}\log\pi_{prior}(L_j) = \log\pi_{prior}(G)$. We set the prior probability of local structure $\pi_{prior}(L_j)$ as $\pi_{prior}(L_j) = \exp\{-(\text{The number of parent genes of } j \text{ th gene})\}$.

By using the dynamic Bayesian network and nonparametric regression model together with the proposed criterion, BNRCd1/namic, we can formulate the network learning process as follows: it is clear from (3) and (6) that the optimization of network structure is equivalent to the choices of

the parent genes that regulate the target genes. However, it is a time-consuming task when we consider all possible gene combinations as the parent genes. Therefore, we cut down the learning space by selecting candidate parent genes. After this step, a greedy hill climbing algorithm is employed for finding better networks. Our algorithm can be expressed as follows:

Step 1: *Preprocessing stage*

We make the p x p matrix whose *(i,* j)th element is the BNRC score of the graph "$\text{gene}_i \rightarrow \text{gene}_j$" and we define the candidate set of parent genes of gene_j that gives small BNRC score. We set the number of elements of the candidate set of parent genes 10.

Step 2: *Learning stage*

For a greedy hill-climbing algorithm, we start form the empty network and repeat the following steps:

Step2-1: For gene_i, implement one from two procedures that *add* a parent gene, *delete* a parent gene, which gives smaller $\text{BNRC}_{dynamic}$ score.

Step2-2: Repeat Step2-1 for prescribed computational order of genes until suitable convergence criterion is satisfied.

Step2-9: Permute the computational order for finding better solution and repeat Step2-1 and 2-2.

Step2-4: We choose the optimal network that gives the smallest $\text{BNRC}_{dyanmic}$ score.

Example 4: Computational Experiment

We demonstrated one embodiment of this invention through the analysis of the *Saccharomyces cerevisiae* cell cycle gene expression data collected by Spellman *et al.* [20]. This data contains two short time series (two time points; cln3, clb2) and four medium time series (18, 24, 17 and 14 time points; alpha, cdc15, cdc28 and elu). In the estimation of a gene network, we used four medium time series. For combining four time series, we ignored the first observation of the target

gene and last one of parent genes for each time series when we fit the nonparametric regression model.

At first, we focused on the cell cycle pathway compiled in KEGG database [22]. The target network is around CDC28 (YBR16Ow; cyclin-dependent protein kinase). This network contains 45 genes and the pathway registered in KEGG is shown in Figure 2 (a) and the estimated network is in Figure 2 (b). The edges in the dotted circles can be considered the correct edges. We thus modeled some correct relations. We denoted the correct estimation by the circle next to edge. The triangle represents the reverse or skip of correct direction. The "x" symbols represent incorrect relationships..

A second example used to demonstrate our methods is the metabolic pathway reported by DeRisi et *al.* [9]. This network contains 57 genes and the target pathway is shown in Figure 3(a).

We applied a Bayesian network and nonparametric regression model [14,15] to this data and the resulting network is depicted in Figure 3 (b). The network of Figure 3 (c) was obtained by the dynamic Bayesian network and nonparametric regression model. It is difficult to estimate the metabolic pathway from cDNA microarray data. However, our model detected correct relationships between the genes. Compared with the Bayesian network and nonparametric regression, the number of false positives of this method depicted in Figure 3 (c) was much smaller than those depicted in Figure 3 (b) by the "x" symbols.

All references cited herein are incorporated herein in their entirety.

References:

1. Akaike, J.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (eds.):2nd International Symposium on Information Theory. Akademiai Kiado, Budapest pp: 267-281 (1973).

2. Berger, J.O.: Statistical Decision theory and Bayesian analysis. Springer-Verlag New York (1985).

3. Bilmes, J.A.: Dynamic Bayesian multinets. Proc. 16th Conference on Uncertainty in

Artificial Intelligence. pp: 38-45 (2000).

4. Burnham, K.P., Anderson, D.R.: Model selection and inference, a practical information-theoretical approach. Springer-Verlag New York (1988).

5. Chen, Tl., He, H.L., Church, G.M.: Modeling gene expression with differential equations. Proc. Pacific Symposium on Biocomputing 4: 29-40 (1999).

6. Davison, A.C.: Approximate predictive likelihood. Biometrika 73: 323-332 (1986).

7. DeBoor, C.: A pracitial guide to splines. Springer-Verlag Berlin (1978).

8. De Hoon, M.J.L., Imoto, S., Kobayashi, K., Ogasawara, NH., Miyano, S.: Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus subtilis* using differential equations. Proc. Pacific Symposium on Biocomputing 8: 2003, in press.

9. DeRisi, J., Lyer, V.R., Brown, P.O.: Exploring the metabolic and gene control of gene expression on a genonmic scale. Science 278: 680-686 (1997).

10. Dierckx, P.: Curve and surface fitting with splines. Oxford (1993).

11. Eiler, P.H.C., Marx, B.: Flexible smoothing with *B*-splines and penalites (with discussion). Statistical Science 11:89-121 (1996).

12. Friedman, N., Murphy, K.., Russell, S.: Learning the structure of dynamic probabilistic networks. Proc. Conf. On Uncertainty in Artificial Ingtelligence pp: 139-147 (1998).

13. Firedman, N., Linial, Ml, Nachman, I., Pe'er, D.: Using Bayesian network to analyze expression data. J. Comp. Biol. 7: 601-620 (2000).

14. Imoto, S., Goto, T., Miyano, Sl.: Estimation of gnetic networks and functional structures between genes by using Bayesian network and noparametric regression. Proc. Pacific Symposium on Biocomputing 7: 175-186 (2002).

15. Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S., Mjiyano, S.: Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. Proc. IEEE Computer Society Bioinformatics Conference; pp: 219-227 (2002).

16. Konishi, S.: Statistical model evaluation and information criteria. In: Ghosh, S. (ed.). Multivariate Analysis, Design of Experiments and Survey Sampling. Marcel Dekker, New

York, pp: 369-399 (1999).

17. Konishi, S., Kitagawa, G.: Generalized information criteria in model selection. Biometrika 83: 875-890 (1996).

18. Pe'er, D., Regev, A., Elidan, G., Friedman, N.: Inferring subnetworks from perturbed expression profiles. Bioinformatics 17: 215-224 (ISBM 2001).

19. Someren, E.V., Wessels, L., Reinders, M.: Linear modeling of genetic networks from experimental data. Bioinformatics 18: 355-366 (ISBM 2002).

20. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cervisiae* by microarray hybridization. Molecular Biology of the Cell 9: 3273-3297 (1998).

21. Tinerey, L., Kadane, J.B.: Accurate approximations for posterior moments and marginal densities. J. Amer. Statist. Assoc. 81: 82-86 (1986).